AI SECURITY – CHEAT SHEET

COMMON AI SECURITY THREATS

Data Poisoning
Injecting malicious data to corrupt model training.

Prompt Injection
Manipulating LLM prompts to bypass instructions.

Model Theft / Extraction
Stealing a model via API queries.

Model Inversion
Reconstructing training data from model outputs.

Membership Inference
Determining if data was in the training set.

AI SECURITY CONTROLS

Input Security
- Input Validation
- Prompt Filtering
- Sanitization

Output Security
- Output Filtering
- Sensitive Data Detection
- Guardrails

Model Security
- Model Validation
- Adversarial Training
- Watermarking

Data Security
- Encryption
- Access Control
- Data Masking

Infrastructure Security
- API Security
- IAM
- Network Security

SECURE AI ARCHITECTURE (BEST PRACTICE)

1. Secure Data Ingestion
2. Model Validation
3. Prompt Filtering
4. API Security
5. Monitoring & Logging

AI SECURITY BEST PRACTICES

- Input Validation
- Output Filtering
- Rate Limiting
- Authentication
- Encryption
- Monitoring & Logging
- Access Control
- Regular Security Testing

MOST ASKED INTERVIEW TOPICS

- Prompt Injection
- Data Poisoning
- Model Theft
- RAG Security
- LLM Security
- AI Governance
- Secure MLOps
- AI Monitoring
- AI Risk Assessment
- AI Red Teaming

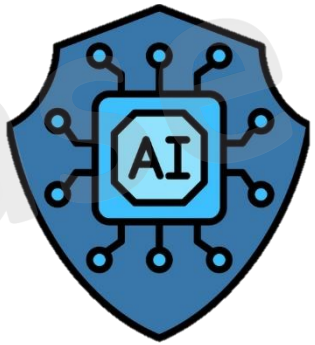
POPULAR AI SECURITY TOOLS

- Microsoft Prompt Shield
- NVIDIA NeMo Guardrails
- OpenAI Guardrails
- Lakera AI
- Protect AI
- HiddenLayer
- Robust Intelligence

Rashmi Bhardwaj

AI MONITORING

- Monitor Query Patterns
- Detect Output Anomalies
- Track Model Performance
- Alert on Suspicious Activity



AI SECURITY FRAMEWORKS

- NIST AI Risk Management Framework
- OWASP Top 10 for LLM Applications
- MITRE ATLAS
- ISO/IEC AI Security Standards

LLM-SPECIFIC RISKS

Prompt Injection
Malicious input overrides system behavior.

Jailbreaking
Bypassing safety guardrails.

Data Leakage
Model reveals sensitive information.

Hallucination Exploitation
Misleading or incorrect outputs used maliciously.

RAG SECURITY RISKS

Document Poisoning
Malicious content in knowledge base.

Data Leakage
Retrieval of sensitive documents.

Unauthorized Access
Improper access to private data.

AI SECURITY TESTING

- **Adversarial Testing** – Test model robustness.
- **Red Teaming** – Simulate real-world attacks.
- **Pen Testing** – Identify vulnerabilities.
- **Drift Testing** – Detect model performance changes.

QUICK REVISION (30-SECOND SUMMARY)

AI Security Protects	Main Attacks	Main Controls
Data, Models, Prompts, APIs, Infrastructure	Prompt Injection, Data Poisoning, Model Theft, Data Leakage	Input Filtering, Output Filtering, Monitoring, Access Control